A Signal Processing View on Packet Sampling and Anomaly Detection

Daniela Brauckhoff*, Kave Salamatian†, Martin May§
*Computer Engineering and Networks Laboratory, ETH Zurich, Zurich, Switzerland
†Computing Department Lancaster University, Lancaster, England
‡Paris Research Lab, Thomson, Paris, France
brauckhoff@tik.ee.ethz.ch, kave@lancaster.ac.uk, martin.may@thomson.net

***Abstract*—Anomaly detection methods typically operate on pre-processed, *i.e.*, sampled and aggregated, traffic traces. Most traffic capturing devices today employ random packet sampling, where each packet is selected with a certain probability, to cope with increasing link speeds. Temporal aggregation, where all packets in a measurement interval are represented by their temporal mean, is then applied to transform the traffic trace to the observation timescale of interest. These pre-processing steps affect the temporal correlation structure of traffic that is used by anomaly detection methods (*e.g.*, Kalman filter, PCA), and have thus an impact on anomaly detection performance. Prior work has analyzed how packet sampling degrades the accuracy of anomaly detection methods; however, neither theoretical explanations nor solutions to the sampling problem have been provided.**

This paper makes the following key contributions: (i) It provides a thorough analysis and quatification of how packet sampling and temporal aggregation modify the signal properties by introducing noise, distortion and aliasing. (ii) We show that aliasing introduced by the aggregation step has the largest impact on the correlation structure. (iii) We further propose to replace the aggregation step with a specifically designed low-pass filter that reduces the aliasing effect. (iv) Finally, we show that with our solution applied, the performance of anomaly detection systems can be considerably improved in the presence of packet sampling.

I. Introduction

A. Motivation

Measuring network traffic is crucial for network operators for the supervision of their networks. Applications using these measurements are, for example, network planning, accounting, and more recently traffic anomaly detection. An important problem with network measurements is related to the burden of capturing, storing, transferring, and processing the huge amount of data generated at the measurement points. Different methods have been proposed to cope with the increasing traffic rates observed in networks. The most prominent of these techniques is packet sampling. Packet sampling is inherently a lossy process, discarding potentially useful information. One has to assess and eventually to compensate for the effects of packet sampling, before using sampled data for networking applications.

The effect of sampling on estimating traffic statistics is a well investigated topic [4]. These studies have shown that packet sampling has indeed an effect on the precision of estimating volume statistics that depends on the sampling rate. Literature has consistently reported that anomaly detection schemes are perturbed by packet sampling, even with relatively high sampling rates [2], [6]. However, no convincing explanations and evaluations have been provided for these observations.

B. A Signal-Processing View on Packet Sampling and Anomaly Detection

Statistical anomaly detection is applied on time series obtained after two consecutive operations: packet sampling and temporal aggregation. Fig. 1 illustrates the different steps that are involved in data preprocessing and anomaly detection. As mentioned before, packet sampling is applied because of router constraints to leverage the burden of packet capture and processing on operational elements in the network. The packet sampling step is generally followed by a temporal aggregation, that consists of summing (or averaging) the amount of data that arrives during a time windows. This step is applied to achieve data compression and to obtain time series at a relevant observation granularity. Statistical anomaly detection methods are applied to the resulting time series. For anomaly detection an entropy reduction step is applied to the data. This entropy reduction generally consits of filtering the normal behavior from the time series. This filtering is typically based on second order statistics that relies on a correct estimation of the temporal correlation structure. The anomaly detection itself is done by detecting rupture in the temporal or spatial correlation structure of the time series obtained after packet sampling and aggregation. These rupture will appear in the filtered signal after entropy reduction. A fairly large spectrum of networking applications falls into this category for example PCA, Kalman filtering, or wavelet-based anomaly detection approaches.

Thus, for analyzing the effect of packet sampling on the performance of statistical anomaly detectors, its impact on the temporal correlation structure needs to be assessed. Fourier theory establishes a strong duality between the frequency and the time domain. Any effect of sampling on the spectra has a (possibly not trivial) effect on the time domain and *vice versa*. Estimating the spectra of traffic can thus provide insight into the effects on anomaly detection. This is the main motivation for taking the detour over spectrum estimation before getting into anomaly detection.

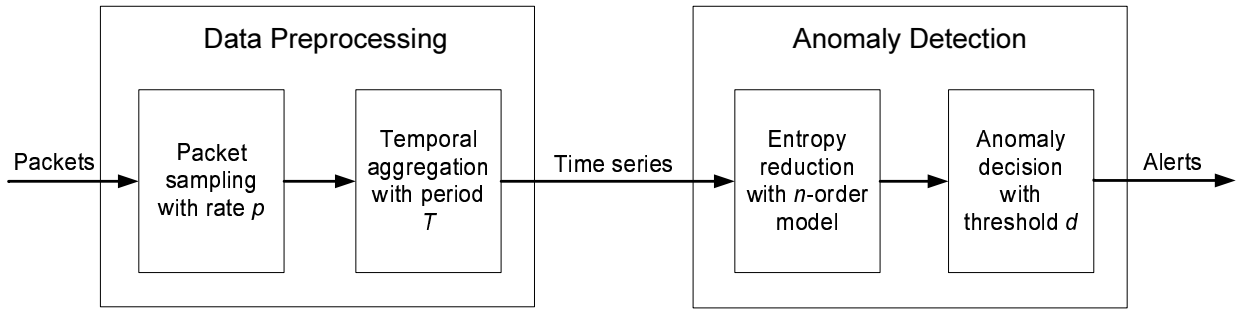

Fig. 1. Block diagram depicting the common-practice steps for pre-processing and anomaly detection. Packets are sampled with a rate p, and then transformed into a time-series by temporal aggregation with period T. Anomaly detection first reduces the time series entropy by applying a model of order n, and then makes an anomaly decision applying a preset threshold d. The output of the anomaly detection systems is a series of alerts.

C. Contributions

In section II we first derive a model for the packet signal at the input of the processing chain. We then carefully study the impact of *packet sampling* and derive the spectrum of the signal at the output of the packet sampling block. We find that packet sampling introduces a wide-band noise that is proportional to the inverse of the packet sampling rate. Next, we examine the impact of *aggregation* and time sampling when applied to the packet sampled signal. We show that this step can be decomposed into an integration (summation) and a regular temporal sampling. We derive the spectrum of the aggregated signal, and show that aggregation and time sampling introduce linear distortion and aliasing.

In section III we advocate an alternative approach to aggregation that applies a specifically designed *low-pass filter* to achieve the same effect as aggregation (*i.e.*, translation to the desired granularity of interest), but without further distorting and aliasing the packet sampled signal. To validate our approach, we compare the spectra of the signal after aggregation and low-pass filtering using synthetic traffic and show the signal-to-noise ratio (SNR) for both variants for real-world traffic. We find that the low-pass filter can effectively avoid the aliasing in the spectra and leads thus to a better SNR estimation.

In section IV we examine the impact of packet sampling and aggregation vs packet sampling and low-pass filtering on the normal behavior modeling assuming an auto-regressive (AR) model is used for calibration and a Kalman filter as whitening filter. We show in particular that the noise of a low-pass filtered signal does not have an important effect on the characteristics of the normal behavior model. Further, we provide a discussion on the impact of both approaches on the detection step.

In section V we validate our findings with real-world data. We show that, when performed correctly, packet sampled data can still be used for anomaly detection. However, there is a fundamental trade-off one has to be aware of: to tackle with the increased noise level introduced by low sampling rates, one has to increase the time scale (or equivalently reduce the bandwidth) of the anomalies to be detected.

We take care throughout the paper to relate signal processing parameters as Signal to Noise Ratio, sampling rate, *etc.*, to parameter relevant to the network practitioner as False alarm rate, detection rate, *etc.* (Section IV). As packet sampling is widely used in practice and we believe the conclusion of this paper to be relevant to practical situations, our ambition is to make this paper accessible to the largest audience of the networking community. We are therefore adding a fair amount of introductory materials in classical sampling theory, that might be seen as trivial by some part of the community but not by all of it. A sign of this last point is that even if applying an anti-aliasing low pass filter before sampling is a trivial step in digital signal processing, we show in the paper that it has been foreseen by the community in the relatively large literature on effect of packet sampling.

II. Sampling and Aggregation

In this section, we investigate the effect of packet sampling and aggregation on the spectrum of the Internet traffic. Therefore we first introduce a model for Internet traffic that is used in the subsequent analysis. We show with the help of spectral analysis that packet sampling is essentially adding a noise to the spectrum. Then we illustrate that aggregation further adds linear distortion and aliasing to the traffic signal.

A. Internet Traffic Model

From an IP layer or above perspective, traffic flowing on a link is a sequence of packets of size L_i arriving at time T_i. Packet arrivals are by nature discrete. However, from the signal processing point of view, an Internet traffic process is a time-continuous process as the arrival time might take any value[1]. Modeling traffic as an evenly-spaced discrete signal of packet sizes means to ignore the arrival time of packets T_i in the analysis and results in dismissal of the temporal context[2].

We model the traffic process as a modulated stochastic point process defined as $X(t) = \sum_i L_i \delta(t - T_i)$, where $\delta(.)$ is the Dirac Delta impulse. We assume that the arrival times T_i and packet sizes L_i are random variables. Generally, this process is assumed to be stationary, *i.e.*, $\mathbb{E}\{X(t)\} = \mu$, $\mathbb{E}\{X(t)X^*(t+\tau)\} = R(\tau)$ and hence its Power Spectral Density (PSD) can be defined.

The process $X(t)$ is built using two ingredients: a continuous point process $\{T_i\}$ defining the packet arrival and

[1]There is indeed a smallest possible time interval due to the speed limitation of the physical link, but we assume it to be very small.

[2]This is applicable in some contexts. See for example [1] where this interpretation is used for applicative flow recognition.

a discrete process $\{L_i\}$ describing the packet size. A closed analytic formula for the PSD is only known for the case where the packet arrival times $S_i = T_i - T_{i-1}$ form a renewal process [10]:

$$\mathcal{S}_X(\Omega) = \lambda\Phi(\Omega) - \lambda^2\mathbb{E}\{L\}^2\delta(\Omega) \quad (1)$$

where λ is the rate of packet arrivals and

$$\Phi(\Omega) = 2\text{Re}\sum_{k\geq 0}\left(\mathbb{E}\{e^{-j2\pi\Omega S}\}\right)^k R_L(k) - R_L(0), \quad (2)$$

where $R_L(k) = \mathbb{E}\{L_i L_{i+k}\}$ is the covariance function of packet sizes and $\mathbb{E}\{e^{-j\Omega S}\}$ is the characteristic function of the distribution of S.

Whenever one knows the distribution of the inter-arrival times S as well as the autocorrelation of the packet sizes $R_L(k)$, one may insert these values into the above formula and derive the PSD analytically. However, as explained before, the formula is only valid for renewal arrivals and no closed formula is known today for more general arrival processes. Unfortunately, empirical observations on Internet traffic are in contradiction with this hypothesis [8]. We have therefore, to resort to a direct estimation of the PSD to rely on as few assumptions as possible.

It is important to note that $\lim_{\Omega\to\infty} = \lambda R_L(0)$ so that the bandwidth of the traffic process becomes infinite. Indeed, the infinite bandwidth is an artifact of the modeling assumption that traffic is a modulated point stochastic process. In fact, at the physical layer a packet is not a Dirac Delta impulse but rather a flat pulse with a duration proportional to the packet length. For example, on a 1 Gbit/s link with a minimal packet size of 60 bytes (40 bytes of TCP/IP header plus 20 bytes of Ethernet header), one would see pulses lasting for 480 nanoseconds, occupying a bandwidth of around 6 MHz. This means that physically speaking the real bandwidth is not infinite, but rather in the order of several megahertz. Our model estimates the spectra correctly up to a bandwidth of several hundred kilohertz. For higher bandwidth, however, one should resort to a precise modeling of the physical layer process. As the bandwidth of interest for anomaly detection is in the order of Hertz, we are save to use the defined model.

B. Impact of Packet Sampling

When packet sampling is applied to a trace, we select a sample of packets to observe, *i.e.* the traffic is only observed at the time of arrival of the selected packets. The PSD of a packet sampled process $\tilde{X}(t)$ can be related to the PSD of the initial process $X(t)$. Let's assume that packet sampling is applied to Internet traffic with intensity λ by keeping each sample with probability Z. The spectrum of the resulting process is obtained as [3]:

$$\mathcal{S}_{\tilde{X}}(\Omega) = \mathbb{E}\{Z\}^2\mathcal{S}_X(\Omega) + \lambda\mathbb{V}\text{ar}\{Z\} \quad (3)$$

This equation shows the effect of packet sampling on PSD estimation. The PSD of the sampled trace consists of (i) the PSD of the initial signal $X(t)$ attenuated by a factor $\mathbb{E}\{Z\}^2$, and (ii) a noise term $\lambda\mathbb{V}\text{ar}\{Z\}$ that translates to a wide-band white noise with variance $\lambda\mathbb{V}\text{ar}\{Z\}$. The signal-to-noise Ratio (SNR) after sampling is equal to:

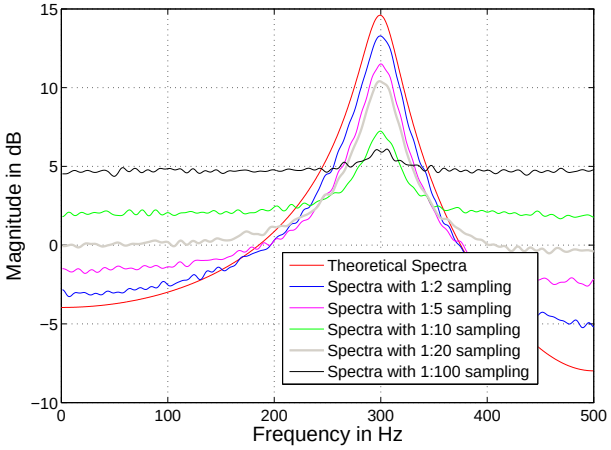

Fig. 2. Power Sprectral Density (PSD) and theoretical spectra for a synthetic packet trace that is sampled with decreasing packet sampling rates. Packet sampling effectively decreases the amplitude of the spectra and increases the noise level.

$$SNR = \frac{\mathbb{E}\{Z\}^2}{B\mathbb{V}\text{ar}\{Z\}}\int_{-\frac{B}{2}}^{\frac{B}{2}}\mathcal{S}_X(\Omega)d\Omega. \quad (4)$$

For a uniform packet selection with probability p the SNR becomes proportional to $\frac{p}{(1-p)}$; for small values of p the SNR becomes approximatively proportional to the sampling rate p.

Let's see if we can reproduce these theoretical results in practice. We first build a suitable synthetic trace using two ingredients: (i) a packet size distribution and (ii) a packet arrival process. Packet sizes are generated as $L_n = L + 100 * l_n$ where L is a fixed value set to $L = 500$ and I_n is an Auto-Regressive (AR) process of order 3 defined as $l_n = \sum_{k=1}^{3} a_k I_{n-k} + \epsilon_n$, with $(a_1, a_2, a_3) = (0.5, 0.6, -0.8)$. The autocorrelation of the packet size can be easily derived numerically using the Wiener-Khinchin theorem as

$$R_L(k) = L^2 + \mathcal{F}^{-1}\left(\frac{1}{|1 + \sum_{k=1}^{3} a_k e^{jk\omega}|^2}\right) \quad (5)$$

where $\mathcal{F}^{-1}(.)$ is the inverse discrete Fourier Transform. The packet arrival is modeled by a renewal process with an exponential distribution of mean $\frac{1}{\lambda}$. We used an arrival rate of $\lambda = 10000$ pkts/sec. The characteristic function of an exponential distribution is given by:

$$\mathbb{E}\{e^{j\Omega S}\} = \frac{j\lambda}{j\lambda + \Omega}. \quad (6)$$

Using the two functions $R_L(k)$ and $\mathbb{E}\{e^{j\Omega S}\}$ and Equation 1, one is able to derive numerically the theoretical form of the PSD for the unsampled signal. The theoretical formula predicting the packet sampled spectra is given in Eq. 3.

To compare the estimated spectra with the theoretically predicted, we applied to the synthetic trace a random packet sampling with different packet sampling rates and derived the PSD with the Capon estimator described in the appendix.

Fig. 2 shows the estimated PSD for the synthetic trace at different sampling rates and compares them with the theoretical spectra [3]. It can be seen that with decreasing sampling rate p, the amplitude of the spectra is reduced and the base level of the spectra changes. For low sampling rates ($p = 0.01$), the spectra totally disappears as it is drowned in the sampling noise.

C. Impact of Aggregation

Aggregation consists of adding up all packets arriving in an interval k of length Ξ and deriving a temporal mean. This translates a packet trace (sampled or not) into a discrete time series $\{\tilde{x}[k]\}$. The motivation for aggregation is two-fold: on one hand practical constraints (as computing power or needed bandwidth to gather the measurements, *etc.*) require data compression; on the other hand the signal gets translated to the desired granularity of interest. For example, in the context of anomaly detection short time scale variations (less than a second) are not really of interest as they could be related to changes in the number of flows sharing a link or to the time dynamic of applications; but variabilities in larger time scales are interesting as they can be related to durable changes such as attacks or failures in equipment.

Aggregation is equivalent to applying to the Internet traffic process an integral operator *i.e.*

$$X^{\Xi}(t) = \frac{1}{\Xi}\int_{t-\Xi}^{t} X(s)ds. \tag{7}$$

followed by a regular temporal time sampling with a period Ξ resulting a discrete signal $\tilde{x}[k]$.

The PSD of a regularly time sampled process $S_Y(\Omega)$ can be expressed using the following well known sampling formula:

$$\mathcal{S}_Y(\Omega) = \frac{1}{\Delta}\sum_{k=-\infty}^{k=+\infty} S_X(\Omega - k\Omega_s) \tag{8}$$

where $\Omega_s = \frac{2\pi}{\Delta}$ is the time sampling frequency in rad/sec. The resulting spectra consists therefore of periodically repeated copies of the Fourier transform of the unsampled signal $X(t)$ that are shifted by integer multiples of the time sampling frequency.

Following the formula for the spectra of regularly time sampled data presented above, and accounting for the rectangular window applied to obtain the aggregation, the PSD of the signal $X^{\Xi}(t)$ resulting from aggregation applied over a process $X(t)$ is given by:

$$\mathcal{S}_{X^{\Xi}}(\Omega) = \frac{1}{\Xi}\sum_{k=-\infty}^{k=+\infty} \text{sinc}^2\left(\frac{(\Omega - k\Omega_{\Xi})\Xi}{2}\right) S_X(\Omega - k\Omega_{\Xi}) \tag{9}$$

where $\text{sinc}(.)$ is the sinc function and $\Omega_{\Xi} = \frac{2\pi}{\Xi}$.

Eq. 9 illustrates two effects of aggregation on the spectra of the process $X(t)$: (i) a linear distortion introduced by the coefficient $\text{sinc}^2(\frac{\Omega\Xi}{2})$; and (ii) a repetition of the PSD of $X(t)$ modulated by the distortion term at regular intervals $\frac{k}{\Xi}$. The linear distortion acts as a non-sharp and non-flat low-pass filter with a 3db cut-off frequency of $f_{\Xi} = \frac{0.44}{\Xi}$. Consequently, by applying aggregation all frequencies larger than f_{Ξ} become highly attenuated. Aliasing happens when S_X has frequency components larger than $\frac{1}{2\Xi}$. Aliasing occurs here with attenuated copies of the spectra. However, as the side lobes are still significant the aliasing effect will be strong, in particular, when high frequencies with large amplitudes exist.

As packet sampling adds a white noise component (see Eq. 3) to the traffic signal, aliasing is very likely to occur when aggregating packet sampled traffic; and this aliasing worsens with lower sampling rates generating higher noise levels. The resulting aggregated process will have properties that are completely different from the initial process, in addition to the dramatic increase of the noise level and a sharp decrease of the SNR. Moreover, if the unsampled signal has frequency components larger than $\frac{1}{2\Xi}$ aliasing will occur even without packet sampling. We will illustrate the effect of aggregation in comparison to our solution in the next section.

III. Solution: Low-pass Filtering

We propose to replace the aggregation block with a specifically designed low-pass filter in order to obtain a better spectrum estimation. The purpose of this filter is to reduce the bandwidth of the signal, such that (i) the bandwidth of interest is still retained and (ii) aliasing is avoided.

A. Impact of Low-pass Filtering

If we assume that the packet sampled signal $\tilde{X}(t)$ has a finite bandwidth B, one can see that if $f_S = \frac{1}{\Delta} \geq 2B$, the shifted replicas resulting from time sampling will not overlap and the resulting discrete PSD will be an exact copy of the initial PSD. This is indeed a re-expression of the Shannon-Nyquist theorem, that states that any band-limited signal $X(t)$ with bandwidth less than B, can be perfectly reconstructed from a sampled sequence $X(k\Delta),\ k = -\infty, \ldots, +\infty$, under the condition that $\frac{1}{\Delta} \geq 2B$. However, if the sampling frequency is too small ($f_S < 2B$), the replicas get mixed and an *aliasing* effect occurs. Aliasing is a major concern with temporal sampling of signals as it means that the PSD at frequency Ω gets garbled with components from frequencies $k\Omega_S - \Omega$. The classical approach to avoid aliasing is to eliminate high-frequency components that lie outside the $[-\frac{f_S}{2}, \frac{f_S}{2}]$ frequency band by applying a low-pass anti-aliasing filter with bandwidth $\frac{f_S}{2}$ before time sampling.

And this is exactly what we are proposing. The packet sampled signal is filtered with a low-pass filter with bandwidth $\frac{1}{2\Xi}$, and afterwards regular time sampling with a sampling rate $f_S > \frac{1}{\Xi}$ is applied. The filtering step has three very important functions: (i) It brings the signal to the relevant granularity of interest by filtering variations with a time scale smaller than 2Ξ. (ii) It acts as an anti-aliasing filter, *i.e.*, it ensures that the following time sampling will not result in aliasing. (iii) As the signal-to-noise ratio depends on the bandwidth as predicted by Eq. 4, it limits the bandwidth and thus the amount of noise that will be introduced in the signal.

By applying this method we ensure that the spectra (and therefore the temporal correlation structure) obtained is exactly

[3]The spectra have been rescaled to the same level of energy.

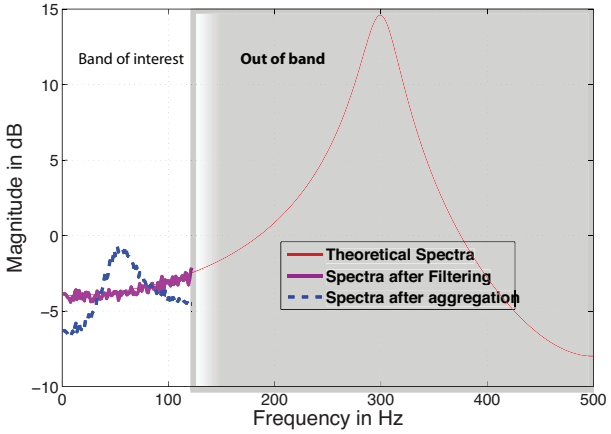

Fig. 3. Power Spectral Density (PSD) estimated over the aggregated and low-pass filtered synthetic trace. The theoretical spectra of the unprocessed signal is added for comparison. The timescale of interest lies between zero and 120 Hz. The PSD of the aggregated signal clearly shows the impact of aliasing, while the PSD of the filtered signal is free of aliasing.

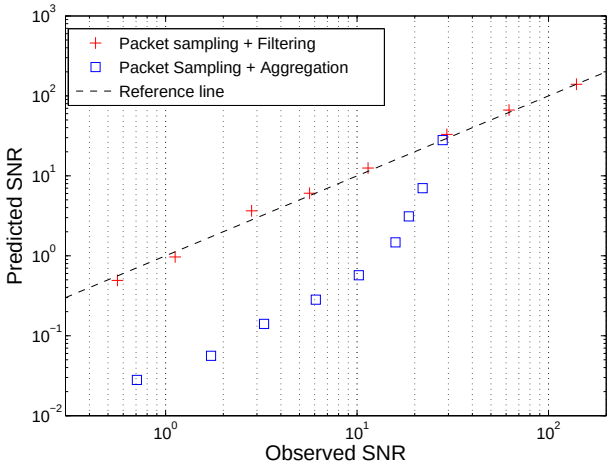

Fig. 4. Empirical SNR vs. theoretical SNR for different sampling rates (each point corresponds to a different sampling rate) for an aggregated and a low-pass filtered real-world packet trace. The empirical SNR of the filtered signal is close to the theoretical SNR (reference line), while the empirical SNR of the aggregated signal is significantly lower than the theoretical SNR for sampling rates smaller than 100%.

the same as the spectra of the initial signal for frequencies below $\frac{1}{2\Xi}$. In fact, the proposed approach replaces the sinc^2 aggregation filter with a better designed low-pass filter that generates less linear distortion in the passband and higher attenuation in the stopband.

B. Aggregation vs. Low-pass Filtering

To illustrate the effect of aggregation and low-pass filtering on the spectra we use again the synthetic trace. In particular, we have applied an aggregation window with a length of 40 μsec to the synthetic trace, resulting in a cut-off frequency of 11 kHz. To obtain the same bandwidth for the low-pass filtered signal, we have designed a filter with a cut-off frequency of 11 kHz, which is followed by a time sampling at 22 kHz. We show the resulting spectra in Fig. 3. The spectra of the aggregated signal shows clearly the effect of aliasing. The peak outside the band of interest (at 300 Hz) generates an artifact (a peak at 52 Hz) inside the band of interest. The spectra of the low-pass filtered signal, on the other hand, almost perfectly estimates the theoretical spectra in the band of interest.

In order to compare aggregation and low-pass filtering in the presence of packet sampling noise, we compute the signal-to-noise ratio after applying each method to a real-world trace from the WIDE project. In particular, we compare the SNR resulting from aggregation with a 1s window (resulting in a cut-off frequency of 0.44 Hz) with the SNR resulting from low-pass filtering with a bandwidth of 0.44 Hz and time sampling with a rate of 1 Hz.

We plot in Fig. 4 the empirical SNR for the aggregated and the low-pass filtered trace vs. the theoretical SNR for a low-pass filtered signal given by

$$SNR = \frac{2\Xi\mathbb{E}\{Z\}^2}{\lambda\mathbb{V}\mathrm{ar}\{Z\}} \int_{-\frac{1}{4\Xi}}^{\frac{1}{4\Xi}} \mathcal{S}_X(\Omega)d\Omega \qquad (10)$$

High sampling rates result in high SNR values and are therefore shown in the right part of the graph. One can see the very good predictive power of the theoretical formula for the SNR after low-pass filtering and sampling. In particular, the proportionality of the SNR with $\frac{p}{1-p}$ is fully validated.

However, what is more instructive is the SNR curve for the aggregated signal. The SNR calculated after aggregation is consistently less than the one for low-pass filtering, especially for low sampling rates and thus small SNR values shown on the left side of the graph. As the noise after packet sampling is the same for aggregation and low-pass filtering, the source of the increased noise level for aggregation is indeed the aliasing effect.

C. Low-pass Filtering in Practice

In practice, low-pass filtering can be efficiently implemented in hardware as an analog filter. Therefore, an analog version of traffic process has to be obtained by applying a digital to analog converter, and the resulting signal has to be converted back with an analog to digital converter. However, if we do not have access to an analog filter, we can still implement the proposed filter in software by using digital signal processing. However, the bandwidth of the initial traffic signal is very large (in the order of several hundreds of megahertz) and the low-pass filter bandwidth will be very small (in the order of hundreds of millihertz). Therefore implementing a low-pass filter with good transition properties in a single step is not possible. Thus, the digital filter implementation consists of a cascade of decimation filters reducing the bandwidth and the sampling rate in several steps.

The complexity involved with low-pass filtering is indeed larger than for aggregation. However, the burden of filtering can be limited by using fewer cascade steps for the decimation filter, and thus trading off the precision of the low-pass filter for reduced complexity. We are pushing this discussion to another paper that will deal with the practical implementation.

IV. Impact on Anomaly Detection

In this section, we illustrate the impact of aggregation vs. low-pass filtering on anomaly detection. In particular, we show

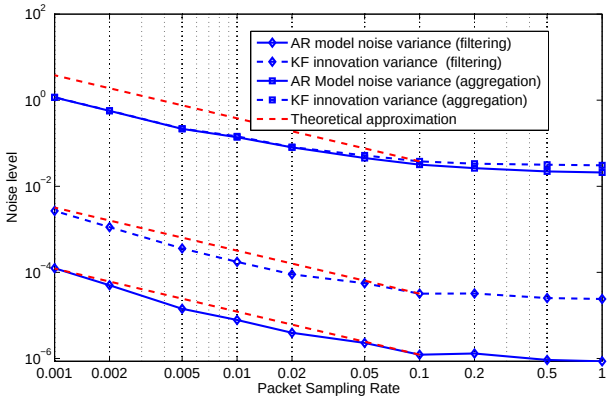

Fig. 5. Comparison of the noise variance captured by an autoregressive (AR) model for a filtered and and aggregated traffic trace as a function of the packet sampling rate. The variance of the innovation process at the output of the Kalman Filter (KF) is also shown. The AR model noise variance for the filtered signal is almost two orders of magnitude smaller than the model noise variance of the aggregated signal.

how the aliasing noise affects the entropy reduction step and the anomaly decision.

A. Entropy Reduction

To illustrate the effect of packet sampling on anomaly detection, we compare here a normal behavior model derived from the signal obtained after aggregation $\bar{x}[k]$, and the signal after low-pass filtering and time sampling $\tilde{x}[k]$. One way to derive a normal behavior model is to use an autoregressive (AR) model [12]. An AR model for the signal $y[k]$ is defined as:

$$y[k] = \sum_{i=1}^{n} \alpha_i y[k-i] + \epsilon_i \tag{11}$$

where ϵ_i is a noise term with variance σ^2. Model calibration consists of choosing the order of the model n, the coefficient α_i and the noise variance σ^2. It is well known that any process can be approximated by an AR model with a high enough order. These parameters can be derived in several ways. We will use in the forthcoming the Burg estimator to estimate the coefficients α_i and σ^2 [13]. The Burg method uses the estimated autocorrelation to derive these parameters. The order of the model is chosen by using a Minimum Description Length criterion trading off the quality improvement resulting from higher order with the increase in the number of parameters [11].

We use the WIDE traces for illustration and set the aggregation window to 1s and the filtering bandwidth to 0.44 Hz (equivalent to the cut-off rate of the aggregation). We obtain an optimal model with an order 5 to 7 AR model in all cases. To enable easier comparison we use for all cases an AR model of order 6.

Let's first analyze the variance of the random term of the model. Intuitively, the noise in the signal input to the modeling phase should be transferred to the model noise, however this relation is not straightforward and no precise relation can be obtained. Equation 3 suggests that the amount of noise resulting from packet sampling will be proportional to $\frac{1-p}{p}$.

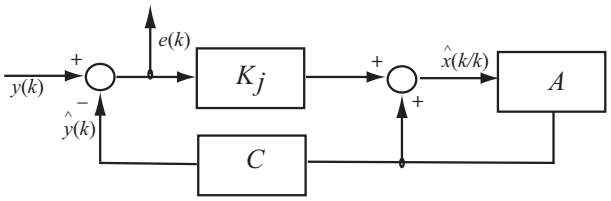

Fig. 7. Block diagram of a Kalman filter.

For small sampling rates p, the noise becomes proportional to $\frac{1}{p}$.

Fig. 5 depicts the AR model error variance for the aggregated signal and the filtered signal, and the expected error variance (proportional to the sampling rate $\frac{1}{p}$). The figure shows that the error variance of the filtered signal is close to the expected error variance at least for sampling rates larger than 1:10. The error variance for the aggregated signal on the other hand overestimates the modeling noise variance by a coefficient of 5. The most striking observation from Fig. 5 is, however, the huge difference of AR model noise variance (almost 2 orders of magnitude) for the filtered and the aggregated traffic signal. We also plot the variance of the innovation process estimated by the Kalman filter as function of the packet sampling rate for the filtered and aggregated signal as we will need this in the following discussion.

The Kalman filter state space model can be stated as

$$\begin{cases} \mathbf{x}[k+1|k+1] = (A - K_\infty CA)\mathbf{X}[k] + K_\infty y[k] \\ \mathbf{e}[k] = -CA\mathbf{x}[k|k] + y[k] \end{cases}$$

where the input is the observed signal $y[k]$, the output is the innovation process $e[k]$ and the state vector is the estimate of the state value. The matrices C, A correspond to the values in the state space representation of the AR model. The transfer function of the Kalman filter can be derived from the above state space representation as

$$W(z) = \frac{1}{1 - CA(zI - A + K_\infty C)^{-1}K_\infty} \tag{12}$$

We show in Fig. 6 the frequency response of the AR model calibrated over the filtered and aggregated signal for unsampled and sampled traffic. One can observe that the AR model transfer function for the filtered signal is not very sensitive to the packet sampling noise. The transfer function obtained from the aggregated signal seems much more sensitive. Last but not least, the graph shows that the filtered signal enables a rich inference of the normal behavior structure, whereas the aggregated signal results in an almost flat spectra. We also plot the transfer function of the Kalman filter for both signals. The figure show clearly the whitening action of the Kalman filter. The transfer function of the Kalman filter approximates very well the inverse of the spectra of the calibrated model, *i.e.*, whenever the Kalman filter is fed with a signal following the spectra of the calibrated AR model, the output will exhibit a flat spectra and will be uncorrelated.

B. Anomaly Decision

The whitening property of the Kalman filter ensures that the innovation signal at the output is an uncorrelated random

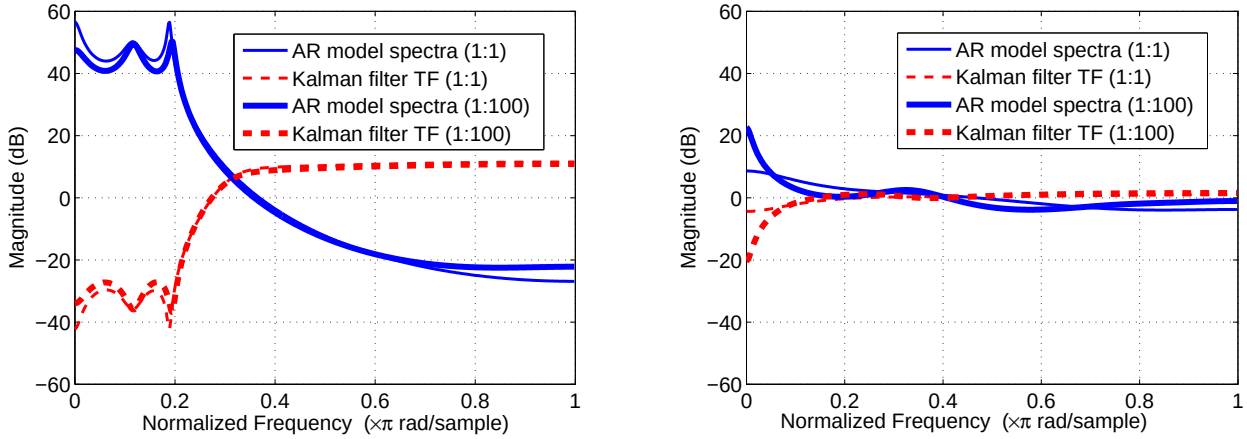

Fig. 6. Spectra of the autoregressive (AR) model and the transfer function (TF) of the Kalman Filter for two signals: low-pass filtered (left) and aggregated (right) at two sampling rates 1:1 and 1:100. The AR model calibrated over the filtered signal is clearly less sensitive to the packet sampling than the model calibrated over the aggregated signal. Further, the filtered signal AR model captures a larger part of the normal behavior structure, while the aggregated AR model spectra is almost flat.

signal with a known variance $V_i(p)$. This variance is a side product of the Kalman filtering algorithm and is used to determine if an observation is normal or not. Let's assume the anomaly signal $a[k]$ and is applied to the system at time 0. One can expect to see along with the anomaly $a[k]$ some normal traffic $n[k]$ crossing the network. So that the traffic entering the Kalman filter is $a[k]+n[k]$. Knowing the transfer function of the Kalman filter (Eq. 12) one can determine exactly how the anomaly will present itself at the output of the whitening filter. From the transfer function, by applying a Z-transform, one can derive the impulse response of the Kalman filter $w[k]$. The output resulting from the anomalous part of the signal $a[k]$ is therefore $e_a[k] = a[k] * w[k] = \sum_{n=0}^{k} a[n-k]w[k]$. Due to the whitening property of the Kalman filter, the signal $e_n[k]$ resulting from the normal part is an uncorrelated signal that can be assimilated to a white noise with a known variance. This means that when an anomaly is present in the traffic, we can see at the output of the anomaly detector a signal $e[k] = e_a[k] + e_n[k]$. An anomaly can be detected only if for some value of k, $e[k] > D$, where D is the anomaly detection threshold. To be on conservative ground, we assume an anomaly signal gets detected when it attains its peak, *i.e.*, $M(a) + e_n[k] > D$.

Under a gaussian assumption for the normal innovation process, [4] the false negative probability P^{MD} can be computed as:

$$P_a^{MD} = 2 * Q(\frac{M(a) - D}{\sqrt{V_i(p)}}) \quad (13)$$

where $V_i(p)$ is the variance of the innovation process at a packet sampling rate p. The coefficient 2 accounts for situations where the maximal amplitude is negative. If the signal is non-gaussian we have to replace the complementary inverse error function $Q(.)$ with the corresponding complementary inverse function of the distribution of the innovation process.

We stated previously that $V_i(p) < \frac{C}{p}$ where C is a constant. Based on Fig. 5 we can set $C = 0.1V_i(0.1)$ as the proportionality assumption holds from $p = 0.1$. This gives $C = 5.7 \times 10^{-4}$ for filtered traffic and $C = 1.9 \times 10^{-2}$ for aggregated traffic (note the coefficient of 34 between these two values). We can therefore write the false negative probability for an anomaly with maximal amplitude after Kalman filtering $M(a)$ as a function of the sampling rate p and the decision threshold D:

$$P_a^{MD}(p, D) < 2 * Q(\sqrt{p}\frac{(M(a) - D)}{\sqrt{C}}) \quad (14)$$

Similarly a false positive occurs when the innovation process value goes beyond the decision threshold when there is no anomaly. The probability of such an event $P_{FA}(p)$ is given by:

$$P^{FA}(p, D) < 2 * Q(\sqrt{p}\frac{(D)}{\sqrt{C}}) \quad (15)$$

$P^{FA}(p, D)$ does not depend on the anomaly and therefore has no subscript. These two values are an upper bound that can be used for design purposes as illustrated later. A ROC curve can be derived by plotting the points $(P^{FA}(D), 1 - P_a^{MD}(p, D))$ for varying values of the decision threshold D.

The derivation presented is related to a single type of anomaly $a[k]$ with maximal amplitude $M(a)$. In practice one will see different types of anomalies with different maximal amplitudes. Let's assume that the distribution of $M(a)$ is given by $P(a)$. Hence, one can expect the overall false negative probability to be bounded by:

$$P^{MD}(p, D) < \int_{a=0}^{\infty} 2 * Q(\sqrt{p}\frac{(M(a) - D)}{\sqrt{C}})P(a)da \quad (16)$$

The overall ROC curve consequently contains the points $(P^{FA}(p, D), P^{MD}(p, D))$ for different threshold values. However, deriving $P(a)$ can be very difficult as we need to have a complete characterization of anomalies. This last point is still a white spot in the research landscape

V. Evaluation

The above analysis gives a precise view on the effect of packet sampling on anomaly detection. Next, we describe how

[4] The Kalman filter as well as PCA based methods are defined in the context of a gaussian hypothesis. They can be used in non-gaussian situations but they will not be anymore optimal.

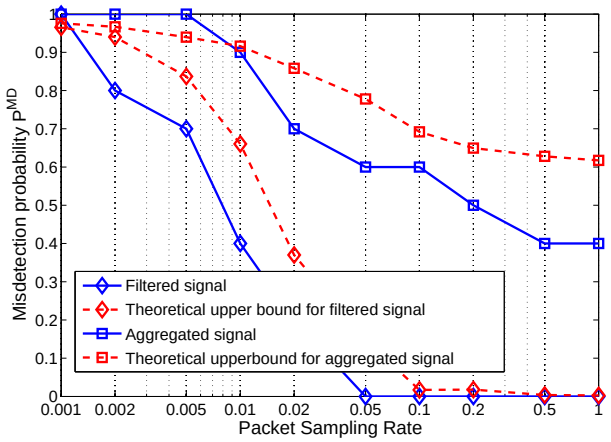

Fig. 8. Comparison of the false negative (misdetection) probability for 20 injected synthetic anomalies with the theoretical upper bound for low-pass filtered and aggregated traffic for different sampling rates. The filtered signal clearly outperforms the aggregated signal.

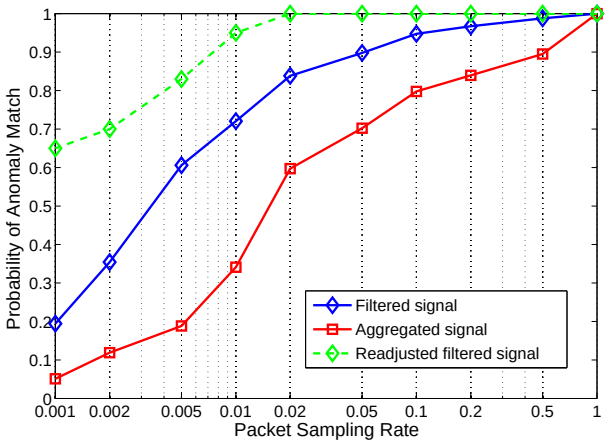

Fig. 9. Comparison of the matching probability obtained over aggregated, low-pass filtered, and readjusted (decreased bandwidth) low-pass filtered traffic signal for different sampling rates. The readjusted filtered signal achieves the highest matching probabilities.

to design an appropriate low-pass filter that has a guaranteed anomaly detection performance at a given sampling rate (that is acceptable for the capturing devices) and for a given anomaly $a[k]$. One can use the two equations 14 (or 16 if the statistics are known) and 15 in order to choose the sampling rate p and the threshold value D. However, this choice might result in a sampling rate that is not acceptable because of router constraints. In this case, one has to select a lower sampling rate and reduce the noise introduced into the signal by adapting the bandwidth of the low-pass filter. This means that we are introducing a trade-off between the bandwidth of interest (the low-pass filter bandwidth) and the packet sampling rate. Anomalies at smaller timescales need higher packet sampling rates.

A. Synthetic Anomalies

Let's first validate the formula for the false negative probability. For this purpose we assume an anomaly that consists of a pulse with duration 5s with an amplitude equal to $0.1\bar{m}$, where $\bar{m}$ is the mean traffic value. According to the frequency response of the Kalman filter given above, we obtain $M(a) = 0.13\bar{m}$ for the low-pass filter and $M(a) = 0.1\bar{m}$ for aggregation. We injected 20 such anomalies in the normal traffic. Further, let's choose a threshold equal to $D = 2.3*\sqrt{V_i(p)}$ as this value gives a $P^{FA}(D) = 0.01$.

We plot in Fig. 8 the false negative (misdetection) probability obtained from the trace for aggregation and low-pass filtering, as well as the theoretical upper bound obtained from Eq. 14. A particularly important observation is that the false negative probability is much larger for the aggregated signal than for the filtered one: all injected anomalies where detected over the filtered traffic up to a sampling rate of 1:20, whereas no more than 60% of the anomalies are detected even over the unsampled aggregated traffic. This was expected as we have shown that the noise introduced by aggregation is much larger than the noise introduced by filtering.

B. Real Traffic Trace

Finally, we validate the findings of this paper by applying all the steps (packet sampling, filtering or aggregation, Kalman filtering and anomaly detection) to a real packet trace. We use a packet trace from an experiment that launched a distributed (from 5 different sources) Denial of Service attack in an operational network. The attacks were generated using the TFN2K tool and consisted of 18 epochs of 100 secs where an attack with an increasing intensity is launched. Each attack is separated by a 300 secs pause period. This experiment was run in the context of the MetroSec project [7] funded by the French government. The attack trace can be obtained upon request from the authors. The nice property of these traces is that they contain known anomalies. However, the complete ground truth cannot be known as one is never sure if there was not an anomaly in the period were no anomaly was detected.

It is noteworthy that the goal of this paper is not to evaluate an anomaly detection method. If this was our goal we would have needed indeed to know as precisely as possible the ground truth. However, the goal of this paper is to evaluate the effect of packet sampling on anomaly detection, *i.e.,* the main performance criteria is the matching probability defined as the likelihood that an anomaly detected on the non-sampled signal is also detected in the sampled signal with the same threshold.

We plot in Fig. 9 the matching probability as a function of the sampling rate for the filtered and aggregated traffic. The plot shows that the anomaly detection performance is less sensitive to packet sampling for the filtered signal than for aggregated signal as the matching probability consistently reaches a larger value for the filtered signal. This observation fully validates (at least on this trace) the proposition of this paper to replace aggregation by low-pass filtering.

To validate the design methodology given above, we have also plotted the matching probability for the readjusted filtered signal. This signal is obtained by decreasing the bandwidth of the low-pass filter by the same coefficient as the packet sampling rate. The basic bandwidth used for the unsampled case is 1 Hz, then if the sampling rate is chosen to be

1:100, we set the bandwidth of the low-pass filter to 0.01 Hz and so on. By doing this we ensure that the increase in input noise resulting from lower packet sampling rates is compensated by a smaller bandwidth. The figure shows that with this readjustment, one can detect the same anomalies than with the unsampled signal up to a sampling rate of 1:50. The performance begins to worsen for larger sampling rate as the decreasing bandwidth of the low-pass filter begins to eliminate some anomalies that have a time scale smaller than 100s. This last observation shows that by using low-pass filtering, we can attain good anomaly detection performance even with high sampling rates on the condition that the bandwidth is reduced accordingly.

VI. Related Work

First and second order network statistics are captured in today's networks for a variety of applications such as accounting, anomaly detection, or network planning. To cope with the increasing packet rates, different sampling methods have been proposed. The two main methods used are systematic sampling where one out of N packets is taken, and random sampling where each packet is taken with a probability of $1/N$.

One line of previous work has concentrated on measuring and quantifying the impact of packet sampling on anomaly detection results. In [2], the authors empirically studied the impact of random packet sampling on volume and distributional anomaly detection metrics. Mai et al. [6] applied a similar methodology for showing the impact of packet sampling on a wavelet-based volume anomaly detection system, and two port scan detection algorithms. Both studies concluded that in general anomaly detection results degrade when N is increased.

A second line of research concerns the question of reconstructing first and second order statistics of interest from sampled traffic views. Duffield et al. [4] have shown how to infer certain first order flow statistics from sampled traffic. Inversion of the flow length distribution from sampled data, which is desirable for monitoring changes in the traffic composition, has also been studied in this context [4], [9], [5].

The only previous work on spectrum estimation from sampled data is that of Hohn and Veitch [5]. The authors provide methods that rely on the theory of point processes for recovering the spectral density of the *aggregated* packet count process when random packet and random flow sampling is applied. They conclude that for larger N (e.g., $N = 1000$) random flow sampling gives still accurate estimates while estimation from packet sampled data is highly inaccurate.

We fill the gap between these two lines of research by studying the impact of packet sampling on the spectral density of the arrival process from a signal processing theory point of view. This allows us to quantify the impact of packet sampling on the spectral density of the arrival process, the aggregated packet count process, and finally on anomaly detection. Moreover, we propose a solution that provides a trade-off between sampling rate and anomaly detection scale.

VII. Conclusion

We have presented an exhaustive discussion on the impact of data pre-processing, namely packet sampling and temporal aggregation, on the performance of anomaly detection systems. We have shown that packet sampling introduces a noise into the anomaly detection signal. We have further shown that popular aggregation techniques add aliasing to the signal.

We proposed to replace the aggregation function with a low-pass filter to prevent the devastating aliasing effects. We evaluated, both theoretically and practically, the effect of signal distortion through packet sampling and aggregation/filtering on the two anomaly detection steps, entropy reduction with a normal behavior model and the subsequent anomaly decision. We evaluated our approach with synthetic anomalies and real traffic traces, and have shown that our filtering solution clearly outperforms temporal aggregation in terms of false positives (misdetection rate) and true positives (detection rate).

References

[1] L. Bernaille, R. Teixeira, and K. Salamatian. Early application identification. In *CoNEXT'06*, December 2006.

[2] D. Brauckhoff, B. Tellenbach, A. Wagner, M. May, and A. Lakhina. Impact of packet sampling on anomaly detection metrics. In *IMC '06*, pages 159–164, New York, NY, USA, 2006. ACM.

[3] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes (2nd ed.)*. Probability and its Applications. Springer, 2003.

[4] N. Duffield, C. Lund, and M. Thorup. Estimating flow distributions from sampled flow statistics. *IEEE/ACM Trans. Netw.*, 13(5):933–946, 2005.

[5] N. Hohn and D. Veitch. Inverting sampled traffic. In *IMC '03*, pages 222–233, New York, NY, USA, 2003. ACM Press.

[6] J. Mai, C.-N. Chuah, A. Sridharan, T. Ye, and H. Zang. Is sampled data sufficient for anomaly detection? In *IMC '06*, pages 165–176, New York, NY, USA, 2006. ACM Press.

[7] Metrosec. The METROSEC project. http://www.laas.fr/METROSEC/.

[8] V. Paxson and S. Floyd. Wide-area traffic: the failure of poisson modeling. In *SIGCOMM '94*, pages 257–268, New York, NY, USA, 1994. ACM.

[9] B. Ribeiro, D. Towsley, T. Ye, and J. Bolot. Fisher information of sampled packets: an application to flow size estimation. In *IMC '06*, pages 15–26, New York, NY, USA, 2006. ACM.

[10] A. Ridolfi. *Power spectra of random spikes and related complex signals*. PhD thesis, EPFL, 2004.

[11] J. Rissanen. *Information and Complexity in Statistical Modeling*. Springer Publishing Company, Incorporated, 2007.

[12] A. Soule, K. Salamatian, and N. Taft. Combining filtering and statistical methods for anomaly detection. In *IMC '05*, 2005.

[13] P. Stoica and R. L. Moses. *Introduction to spectral analysis*. 1997.